PNeSM: Arbitrary 3D Scene Stylization via Prompt-Based Neural Style Mapping

Jiafu Chen[1], Wei Xing[1*], Jiakai Sun[1], Tianyi Chu[1], Yiling Huang[1]
Boyan Ji[1], Lei Zhao[1*], Huaizhong Lin[1*], Haibo Chen[2], Zhizhong Wang[1*]
[1]Zhejiang University
[2]Nanjing University of Science and Technology
{chenjiafu, wxing, csjk, chutianyi, ji_by, cszhl, linhz, endywon}@zju.edu.cn,
huangyiling@hotmail.com, hbchen@njust.edu.cn

Abstract

3D scene stylization refers to transform the appearance of a 3D scene to match a given style image, ensuring that images rendered from different viewpoints exhibit the same style as the given style image, while maintaining the 3D consistency of the stylized scene. Several existing methods have obtained impressive results in stylizing 3D scenes. However, the models proposed by these methods need to be re-trained when applied to a new scene. In other words, their models are coupled with a specific scene and cannot adapt to arbitrary other scenes. To address this issue, we propose a novel 3D scene stylization framework to transfer an arbitrary style to an arbitrary scene, without any style-related or scene-related re-training. Concretely, we first map the appearance of the 3D scene into a 2D style pattern space, which realizes complete disentanglement of the geometry and appearance of the 3D scene and makes our model be generalized to arbitrary 3D scenes. Then we stylize the appearance of the 3D scene in the 2D style pattern space via a prompt-based 2D stylization algorithm. Experimental results demonstrate that our proposed framework is superior to SOTA methods in both visual quality and generalization.

Introduction

3D scene stylization is an important editing task in vision and graphics, which facilitates the creation of new artistic scenes. Given a 3D scene and a style image, 3D scene stylization models can generate stylized images of the scene from arbitrary novel views. Naively applying approaches designed for image/video stylization to 3D scenes often leads to inconsistent results due to the lack of 3D information. To handle the inconsistency problem, several methods (Huang et al. 2021; Höllein, Johnson, and Nießner 2022; Cao et al. 2020; Kopanas et al. 2021) have explored 3D scene stylization based on explicit representations (*e.g.*, meshes, voxels and point clouds). However, their discrete representation of scenes will lead to a loss of precision in geometry.

Recently, Neural Radiance Field (NeRF) (Mildenhall et al. 2020) proposes to use neural networks for continuous scene modelling. Due to its excellent performance in reconstructing both geometry and appearance, Stylizing-3D-scene (Chiang et al. 2022) first introduces NeRF for 3D scene stylization. It fixes the geometry of the scene and changes only the appearance by using a hypernetwork to predict parameters for calculating artistic appearance. To further improve the visual quality of stylization, several approaches (Huang et al. 2022; Zhang et al. 2022; Nguyen-Phuoc, Liu, and Xiao 2022; Fan et al. 2022; Liu et al. 2023) have been developed. StylizedNeRF (Huang et al. 2022) and SNeRF (Nguyen-Phuoc, Liu, and Xiao 2022) are proposed to mutually optimize image stylization module and scene appearance representation to fuse the stylization ability of 2D stylization network with the 3D consistency provided by NeRF. ARF (Zhang et al. 2022) minimizes the distance between each feature vector in an image rendered from NeRF and its nearest neighbor feature vector in the given style image. INS (Fan et al. 2022) simultaneously changes both the geometry and appearance to enable a more flexible stylization by stylizing shape tweaks on the scene surface. StyleRF (Liu et al. 2023) transforms the grid features of the scene according to the reference style. However, despite valuable efforts, these methods still entangle geometry and appearance to some extent, necessitating re-training for a new scene.

In this work, we propose the **first** framework for arbitrary 3D scene stylization, *which can not only transfer arbitrary styles but also stylize arbitrary 3D scenes with only one stylization model*. The key insight is a Prompt-based Neural Style Mapping (PNeSM) which disentangles the geometry and appearance of a 3D scene by mapping the appearance into a 2D style pattern space and then stylizes the appearance in the 2D style pattern space via a prompt-based 2D stylization algorithm.

3D scene disentanglement consists of two main parts: UV mapping and appearance mapping. Inspired by (Xiang et al. 2021), *the UV mapping* trains a UV mapping network to project the 3D real-world coordinates into a 2D (UV) style pattern space. Different from (Xiang et al. 2021), we use voxel-grid representation instead of MLP for fast training and use a stylization network rather than manual operation to change the appearance. Thanks to the complete separation of geometry and appearance, the stylization can be conducted in a unified style pattern space. *The appearance mapping* reconstructs the original appearance of the scene, which maps the projected style pattern coordinate to the radiance color through an MLP.

*Corresponding Authors.

3D scene stylization is realized via prompt-based stylization mapping. *The prompt-based stylization mapping* stylizes the appearance of the scene in the 2D style pattern space. Given arbitrary style images, a powerful pre-trained 2D stylization network (*e.g.*, SANet (Park and Lee 2019)) can generate their corresponding 2D style patterns. However, directly using these 2D style patterns to stylize the appearance of 3D scenes would easily lead to disorganized results (as shown in Fig. 6). This is because they are generated and applied to the appearance of 3D scenes without taking the geometry information into consideration. To address this problem, we integrate a visual prompt to the feature maps of the bottleneck layer of the pre-trained 2D stylization network, and it is the only tensor we need to train in the stylization stage. When trained on a single 3D scene, the prompt can be treated as a scene-related adaptor, adapting the 2D style patterns to be aware of the specific geometry information of that scene. When trained on multiple 3D scenes, it can be treated as a scene-agnostic adaptor, adapting the 2D style patterns to be aware of the universal geometry information which is tolerant of diverse geometric variations. In our experiments, we find the prompt can generalize well to unseen scenes by learning on just few-shot (*e.g.*, 3) scenes. The scene-agnostic prompt thereby enables our framework to achieve arbitrary 3D scene stylization.

We have conducted comprehensive experiments to demonstrate the effectiveness and superiority of our proposed method. Experimental results demonstrate that our method not only achieves high-quality 3D scene stylization, but also generalizes well to unseen styles and unseen scenes.

Overall, the contributions can be summarized as follows:

- We propose a novel 3D scene stylization framework, *i.e.*, PNeSM, which realizes complete disentanglement of the geometry and appearance of 3D scenes by mapping the appearance of 3D scenes into a 2D style pattern space. The stylization of 3D scenes is carried out in an independent and unified 2D style pattern space, which allows our framework to be generalized to any 3D scene. For each new 3D scene, there is no need to train a separate stylized model.
- To the best of our knowledge, we are the first to explore the use of prompt learning to adapt the pre-trained 2D stylization network for 3D scene stylization, which provides a simple yet effective way to improve the quality of stylized 3D scenes.
- Extensive experiments on different datasets are conducted to demonstrate the effectiveness and superiority of our method in visual quality when generalizing to new 3D scenes and new styles.

Related Work

Image/Video Style Transfer

Image style transfer aims to create new artworks from real-world photos by using style information from real artworks. (Gatys, Ecker, and Bethge 2016) proposed an optimization-based style transfer method. However, the iterative optimization process is prohibitively slow. Motivated by this, several approaches (Johnson, Alahi, and Fei-Fei 2016; Li and Wand 2016; Ulyanov et al. 2016; Huang and Belongie 2017; Li et al. 2017; Park and Lee 2019; Liu et al. 2021a) have been developed based on feedforward networks. With such rapid progress, satisfying artistic image can be easily generated.

Video style transfer (Gao et al. 2018; Chen et al. 2017; Deng et al. 2021; Wang et al. 2020) takes on the challenge of maintaining the consistency between adjacent frames in the stylized video, eliminating flickering effects. This is achieved by introducing optical flow or aligning intermediate feature to constrain nearby video frames.

Since both image and video style transfer can only stylize 2D images, lacking the knowledge of 3D scene, simply applying them to 3D scene stylization often leads to inconsistency between different views.

3D Scene Style Transfer

3D scene style transfer requires that the images rendered from arbitrary viewpoints of the stylized scene match the style reference. Several approaches (Huang et al. 2021; Höllein, Johnson, and Nießner 2022; Cao et al. 2020; Kopanas et al. 2021; Mu et al. 2022) have been developed using explicit 3D models (*e.g.*, meshes, voxels and point clouds). For example, (Huang et al. 2021) modulates scene features in point cloud with the given style image. However, these methods are limited by their quality of geometry reconstruction.

To offer a more faithful representation of scenes, some researchers turn to performing style transfer on radiance field (Chiang et al. 2022; Huang et al. 2022; Zhang et al. 2022; Nguyen-Phuoc, Liu, and Xiao 2022; Fan et al. 2022; Liu et al. 2023; Chen et al. 2023). Stylizing-3D-Scene (Chiang et al. 2022) is the first to introduce NeRF to 3D scene style transfer, conducting patch-based optimization on content and style losses. StylizedNeRF (Huang et al. 2022) and SNeRF (Nguyen-Phuoc, Liu, and Xiao 2022) respectively propose mutual learning and alternate training strategy to effectively reduce GPU memory requirements. ARF (Zhang et al. 2022) explores improving style details on stylized renderings with single style image and proposes a deferred back-propagation strategy to directly optimize on full-resolution images. INS (Fan et al. 2022) proposes a method to interpolate between different styles in its pre-defined set and generates renderings by the new mixed styles. StyleRF (Liu et al. 2023) conducts 3D scene stylization within the feature space of a radiance field and designs sampling-invariant content transformation to maintain multiview consistency. All of these methods require re-training a stylized model for every unseen scene. In this paper, our method focusses on not only transferring arbitrary styles, but also stylizing arbitrary 3D scenes with only one stylized model.

Prompt Learning

Prompt learning has emerged as a prominent technique in natural language process (NLP), with the hope to adapt pre-trained large language models, which are frozen, to downstream tasks by reformulating their input text. Building on domain-specific knowledge, some works (Brown et al. 2020;

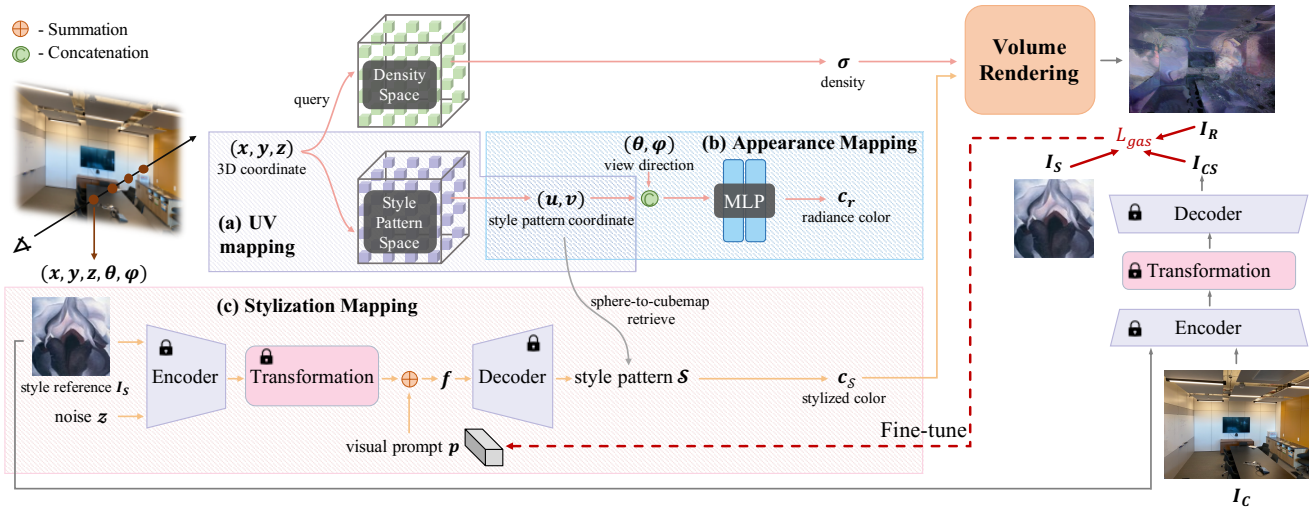

Figure 1: An overview of our method. (a) UV mapping is designed to support the complete disentanglement of geometry and appearance, which maps a 3D coordinate to a style pattern coordinate. (b) To reconstruct the original appearance, we use appearance mapping to map the style pattern coordinate along with the view direction to the radiance color $\mathbf{c_r}$. (c) A pre-trained image stylization network integrated with a visual prompt is used for stylization mapping, stylizing the appearance of the scene in the 2D style pattern space.

Cui et al. 2021; Petroni et al. 2019) manually design text prompts, achieving impressive results in few-shot or even zero-shot settings. To further unleash prompt's power, recent works propose to treat the prompt as task-specific variable and optimize it via backpropagation, namely Prompt Tuning (Lester, Al-Rfou, and Constant 2021; Li and Liang 2021; Liu et al. 2021b; Zhong, Friedman, and Chen 2021).

Inspired by the success of prompt learning in NLP, researchers begin experimenting with applying prompts to computer vision. (Zhou et al. 2022b) and (Zhou et al. 2022a) transform context words into a set of learnable vectors for downstream image recognition, introducing prompt learning to vision-language models. VPT (Jia et al. 2022) and VP (Bahng et al. 2022) explore prompting with images. VPT prepends a set of tunable parameters to ViT (Dosovitskiy et al. 2021) in each Transformer encoder layer, which outperforms full fine-tuning ViT in many cases and reduces per-task storage cost. VP directly adds prompt as perturbations to the image in pixel space. These approaches show that prompt learning has great potential in visual domain. Motivated by visual prompt, we introduce prompt learning in feature level to style transfer field, adapting the pre-trained 2D image style transfer network for 3D scene stylization.

Proposed Method

As illustrated in Fig. 1, our proposed *Prompt-based Neural Style Mapping (PNeSM)* consists of three main parts: (a) A UV mapping that projects the 3D real-world coordinates into a 2D (UV) style pattern space, disentangling appearance from geometry. (b) An appearance mapping that maps the UV style pattern coordinate to the radiance color, representing the original appearance of the scene. (c) A prompt-based stylization mapping that stylizes the appearance of the scene in the 2D style pattern space, obtaining the final stylized color. To train PNeSM, we exploit a two-stage training strategy: I) a disentanglement stage which jointly trains the UV mapping and appearance mapping to reconstruct the scene, and II) a stylization stage which only trains the prompt-based stylization mapping for scene stylization.

In the following subsections, we first provide a thorough review of our scene representation, NeRF, as preliminary. Then, we introduce how to completely disentangle appearance from geometry at the disentanglement stage. Finally, we introduce how to achieve stylization on the appearance at the stylization stage.

Preliminary

NeRF (Mildenhall et al. 2020) proposes to encode a 3D scene as a function, $f : (x, y, z, \theta, \phi) \rightarrow (\sigma, \mathbf{c_r})$, which maps a 3D coordinate (x, y, z) and its view direction (θ, ϕ) to a volume density σ and a radiance color $\mathbf{c_r}$.

During volume rendering, rays $\mathbf{r}$ casting from the camera pass through the pixel of captured images. The pixel color thus can be calculated by sampling N points between t_n and t_f (the near and far bound):

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))\mathbf{c_r}_i, \quad \text{where } T_i = \exp(-\sum_{j=1}^{i-1}\sigma_j\delta_j), \tag{1}$$

where $\delta_i = t_{i+1} - t_i$ denotes the distance between adjacent samples.

Given training images, NeRF model is optimized by minimizing the L_2 distance between the observed pixel $C(\mathbf{r})$ and

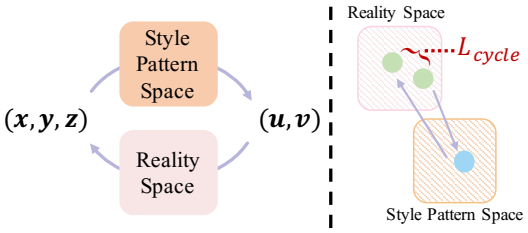

Figure 2: With the cycle loss L_{cycle}, we encourage the bijective mapping between real-world 3D coordinate and 2D style pattern coordinate.

the rendered pixel $\hat{C}(\mathbf{r})$:

$$L_{rec} = \sum_{\mathbf{r}\in\mathcal{R}} \| \hat{C}(\mathbf{r}) - C(\mathbf{r}) \|_2, \tag{2}$$

where $\mathcal{R}$ is a ray batch from training views.

Appearance-Geometry Disentanglement

NeRF models radiance color (*i.e.*, scene appearance) using 3D coordinates and view directions as input. However, it entangles appearance and geometry in a "black-box" that cannot be edited. The target of 3D scene stylization is to stylize the appearance of the scene while retaining its geometry. Therefore, a critical desideratum is to disentangle appearance from geometry. Inspired by Neural Texture Mapping (NeuTex) (Xiang et al. 2021), we add a UV mapping network to explicitly disentangle appearance from geometry by projecting real-world 3D coordinates into a 2D (UV) style pattern space during disentanglement stage. In this way, we can obtain the color of points on rays through their mapped style pattern coordinates, thereby enabling stylization of the scene's appearance in the unified 2D style pattern space. Each scene shares the same style pattern space. The disentanglement is achieved during conducting simple reconstruction training, where an appearance mapping (an MLP) is used to map the UV style pattern coordinate, along with view direction, to the radiance color $\mathbf{c_r}$, which represents the original appearance of the scene. After disentanglement, each UV style pattern coordinate in the style pattern space can pinpoint a specific point in the 2D style pattern via a sphere-to-cubemap retrieval operation (explained further in supp.). To speed up training, we use voxel-grid representation (Sun, Sun, and Chen 2022) instead of MLP (Xiang et al. 2021) to model our UV mapping network and the modality of density.

Following (Xiang et al. 2021), we also employ a cycle loss to ensure the rationality of the style pattern space, avoiding mapping multiple points in reality space to the same point in style pattern space. As shown in Fig. 2, we train another inverse mapping network to project the style pattern coordinate to reality space. In particular, for each ray, we focus more on whether the sample points that significantly contribute to the final pixel color maintain an accurate cycle mapping, reflecting the surface of the scene. From Eq. (1), the contribution of each point is evident, so we consider it as the weight:

$$w_i = T_i(1 - \exp(-\sigma_i\delta_i)). \tag{3}$$

The cycle mapping process and cycle loss are depicted in Fig. 2 and defined as:

$$(x, y, z) \rightarrow (u, v) \rightarrow (x', y', z'), \tag{4}$$

$$L_{cycle} = \sum_i w_i \|(x, y, z) - (x', y', z')\|_2. \tag{5}$$

The full loss function L for scene reconstruction is:

$$L = \lambda_{rec}L_{rec} + \lambda_{cycle}L_{cycle}. \tag{6}$$

Note that though our appearance-geometry disentanglement of NeRF is based on NeuTex (Xiang et al. 2021), there are two key differences: (1) The UV mapping network is formulated by voxel-grid representation (Sun, Sun, and Chen 2022) instead of MLP (Xiang et al. 2021), which can greatly speed up training while maintaining the reconstruction quality. (2) We add a new prompt-based stylization mapping to stylize the appearance of the scene in the 2D style pattern space, which is aware of the geometry information of the scene and can stylize the appearance more harmoniously.

Prompt-based Appearance Stylization

After the disentanglement stage, we can intuitively stylize the appearance of the scene in the unified 2D style pattern space via stylization mapping. We first generate a new 2D style pattern $\mathcal{S}$ given a reference style image I_S. Next, we locate each UV coordinate in the style pattern space to the corresponding pixel in the 2D style pattern via sphere-to-cubemap retrieval. Specifically, for a UV coordinate (u, v), we retrieve the pixel $\mathbf{c}_\mathcal{S}$ in the 2D style pattern. $\mathbf{c}_\mathcal{S}$ is the stylized color for (u, v), signifying the newly stylized appearance of the scene.

As demonstrated in (Gatys, Ecker, and Bethge 2016; Li et al. 2017), pre-trained 2D stylization approaches are effective in extracting the texture information from style images. To learn only style patterns and remove contents in the style image, a noise image $\boldsymbol{z}$ is utilized as the content input. Subsequently, the style patterns are employed to change the appearance of the scene within the style pattern space. However, due to the lack of consideration on scene's geometry, directly using the style patterns explained above would easily lead to disorganized results, as will be demonstrated later in Fig. 6. A satisfactory stylized scene should not only exhibit pleasant style patterns in the appearance, but also harmoniously fuse the style patterns with the scene geometry inherently. Therefore, the geometric awareness is important for 3D scene stylization and must be properly considered.

In order to integrate geometric information into the generated style patterns, we can fine-tune the decoder of a pre-trained 2D stylization network under the supervision of a

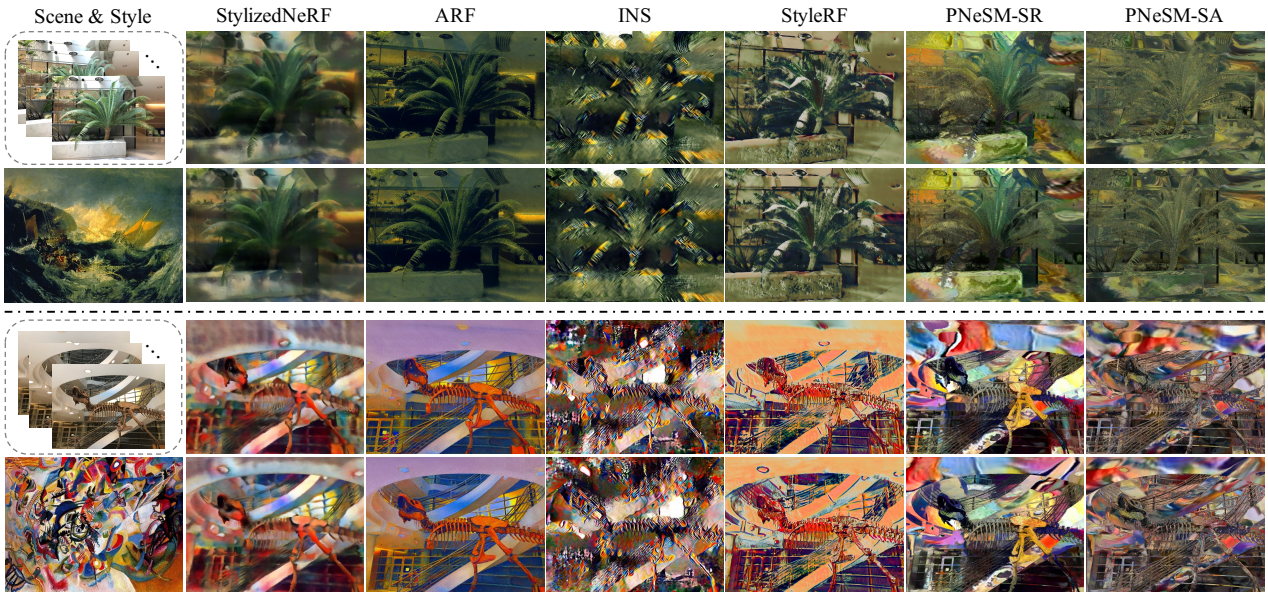

Figure 3: Qualitative comparisons on LLFF dataset. We compare our method to StylizedNeRF (Huang et al. 2022), ARF (Zhang et al. 2022), INS (Fan et al. 2022) and StyleRF (Liu et al. 2023). Our method stylizes scenes with clear geometry and competitive stylization quality.

geometry-aware stylization loss L_{gas}:

$$\begin{aligned} L_{gas} &= \sum \parallel I_{CS} - I_R(\boldsymbol{\theta}) \parallel_2 \\ +\lambda_{style} &\sum_i \|\mu(\phi_i(I_S)) - \mu(\phi_i(I_R(\boldsymbol{\theta})))\|_2 \\ +\lambda_{style} &\sum_i \|s(\phi_i(I_S)) - s(\phi_i(I_R(\boldsymbol{\theta})))\|_2, \\ \boldsymbol{\theta}^* &= \arg\min_{\boldsymbol{\theta}} L_{gas}, \end{aligned} \tag{7}$$

where $I_R(\boldsymbol{\theta})$ denotes the rendered image given the fine-tuned 2D stylization network $\boldsymbol{\theta}$ and I_{CS} denotes the stylized image using training views as content inputs to pre-trained 2D stylization network. μ and s are channel-wise mean and standard deviation, respectively. ϕ_i denotes a layer in VGG-19. The first term aligns stylized training views and rendered views from the scene, thus the style patterns generated by the 2D stylization network should be aware of the geometry information of the scene. The last two terms calculate the style loss between rendered images and the style reference, in the manner typically employed in image stylization methods.

However, fine-tuning the decoder of a pre-trained 2D stylization network is cumbersome and time-consuming, making it inflexible in practical. To alleviate this problem, we introduce prompt learning for fast and flexible adaptation. To be specific, we add a visual prompt $\boldsymbol{p}$ to the output feature maps from the style transformation module of the 2D stylization network. The visual prompt is treated as an extra and independent learnable component implicitly representing geometry information of scenes. We train the visual prompt using L_{gas} to generate more harmonious style pattern for scenes during stylization. All parameters of image stylization network are frozen, and the visual prompt is the only parameter requires training at the stylization stage. This means that $\boldsymbol{\theta}$ in Eq. 7 corresponds to $\boldsymbol{p}$ in our method. Note that our method is not limited to a specific 2D stylization network and the ability to transfer arbitrary styles is inherently embedded within the arbitrary image stylization network. The visual prompt is plug-and-play and can be easily integrated into existing image style transfer methods.

Experiments

Implementation Details

We implement our model during the disentanglement stage on top of DVGO (Sun, Sun, and Chen 2022), where we replace the feature grid as described in (Sun, Sun, and Chen 2022) with a grid designed for style pattern space. Following (Sun, Sun, and Chen 2022), we use the Adam optimizer with a learning rate of 0.1 for all voxels and 0.001 for MLP. λ_{rec} and λ_{cycle} are set to 1. During stylization, we adopt SANet (Park and Lee 2019) as the image style transfer network. The visual prompt is trained for 5k iterations using an Adam optimizer with a learning rate of 0.1. λ_{style} is set to 0.1. We use relu1_1, relu2_1, relu3_1, and relu4_1 layers in VGG-19 to calculate loss in Eq. 7. Appearance-geometry disentanglement is scene-related, while prompt-based appearance stylization is scene-agnostic. For arbitrary test scenes, their appearance and geometry should be disentangled first, and then the stylization can be conducted by the stylization mapping module trained on training scenes. All experiments are performed on a single NVIDIA RTX A6000 (48G) GPU.

Datasets. Following previous image stylization methods, we take WikiArt (Karayev et al. 2013) as the style dataset. We conduct extensive experiments on real-world scenes,

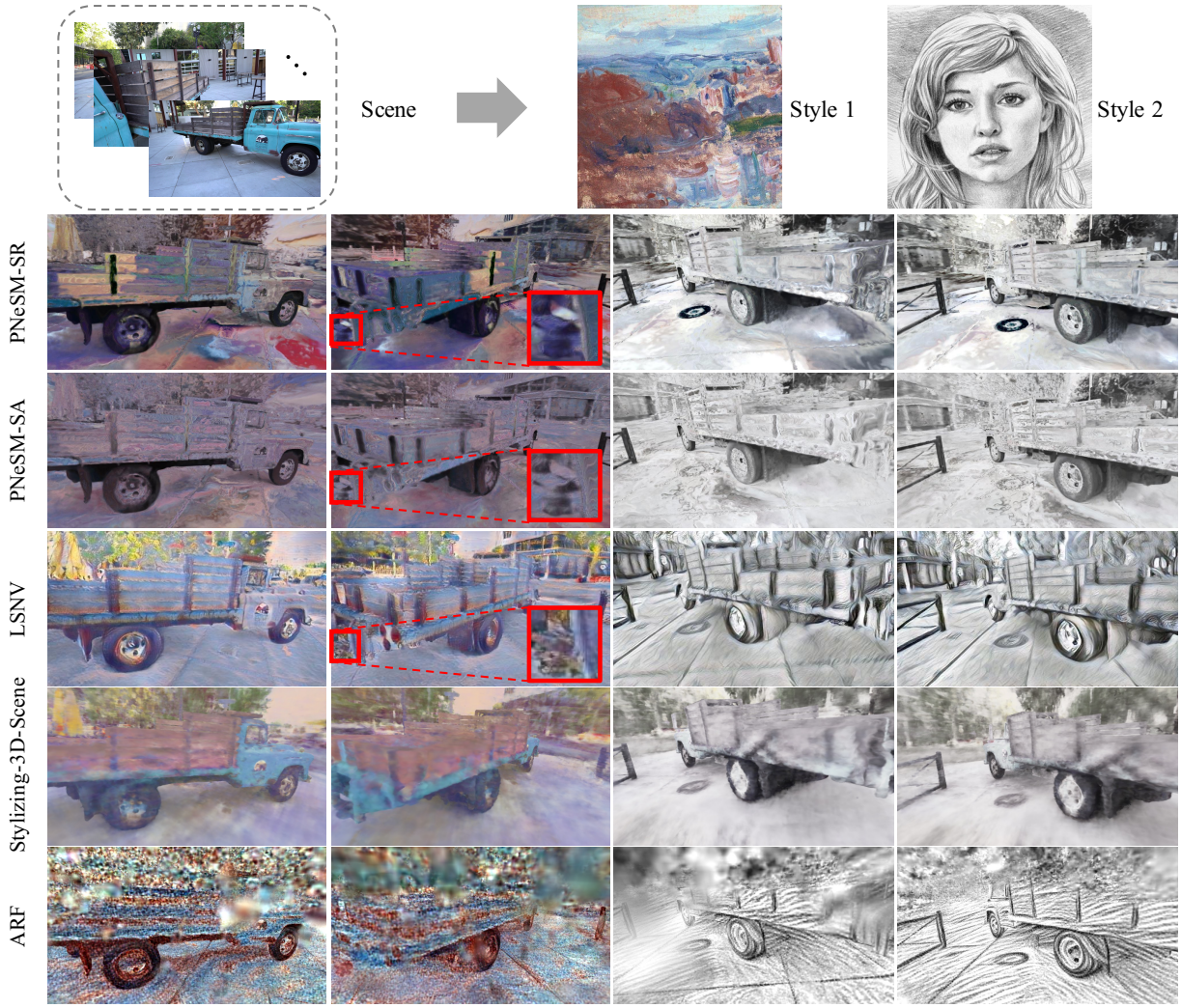

Figure 4: Qualitative comparisons on Tanks and Temples dataset. We compare our method to LSNV (Huang et al. 2021), Stylizing-3D-Scene (Chiang et al. 2022) and ARF (Zhang et al. 2022). Stylized scenes generated by our method contain both precise geometry and pleasant stylization.

forward-facing LLFF (Mildenhall et al. 2019) and 360° unbounded Tanks and Temples dataset (Knapitsch et al. 2017). The training sets of LLFF dataset are *Room, Horns, Leaves, Flower, Orchids*, and we use *Fern, Trex* for evaluation. On Tanks and Temples dataset, we use *Playground, Horse, Francis* for training, and evaluate on *Truck*.

Baselines. On LLFF dataset, we compare our method to StylizedNeRF (Huang et al. 2022), ARF (Zhang et al. 2022), INS (Fan et al. 2022) and StyleRF (Liu et al. 2023). On Tanks and Temples Dataset, we compare our method to LSNV (Huang et al. 2021), Stylizing-3D-Scene (Chiang et al. 2022) and ARF (Zhang et al. 2022). For all these methods, we use their released codes and pre-trained models. Among them, LSNV is based on point cloud scene representation, while others are based on NeRF. We do not conduct comparison on image/video style transfer methods, which are less competitive than 3D scene stylization approaches proven in previous works (Huang et al. 2021, 2022; Nguyen-Phuoc, Liu, and Xiao 2022; Chiang et al. 2022).

Qualitative Results

We experiment with both scene-related (PNeSM-SR) and scene-agnostic (PNeSM-SA) visual prompt on our method.

LLFF. In Fig. 3, we show qualitative comparisons on LLFF dataset. We observe that StylizedNeRF (Huang et al. 2022) degrades the scene in clarity, which might be caused by introducing spatial consistency to 2D stylization network and training style module for NeRF with the supervision of fine-tuned 2D stylization results. ARF (Zhang et al. 2022) sometimes produces plain results in the aspect of color tone (*e.g.* 3rd and 4th rows). INS (Fan et al. 2022) disrupts the geometry of scenes, yielding poor-quality stylizations. StyleRF (Liu et al. 2023) shows low similarity between stylized scenes and style images. In contrast, our method can not only maintain clear geometry, but also change the appearance of the scene resembling the reference style. Our method

Methods	StylizedNeRF	ARF	INS	StyleRF	PNeSM
Short-range	0.0229	0.0125	0.0208	0.0235	0.0116
Long-range	0.0627	0.0353	0.0439	0.0531	0.0351

Table 1: Short-range and Long-range consistency comparison. The lower the better.

shows better stylization quality in terms of style transformation. (Please refer to the quantitative comparison on *style loss* in supp.)

Tanks and Temples. In Fig. 4, we qualitatively compare our results with baselines on Tanks and Temples dataset. LSNV (Huang et al. 2021) reconstructs the scene with point cloud, whose geometry is not precise and further damages the stylization result. Stylizing-3D-Scene (Chiang et al. 2022) calculates Gram matrix loss (Gatys, Ecker, and Bethge 2016) on sub-sampled patches to achieve stylization. Due to the limited receptive field, the stylized results are blurry and the stylization quality is poor. ARF (Zhang et al. 2022) contains geometry artifacts for Tanks and Temples dataset on their implementation based on Plenoxel (Fridovich-Keil et al. 2022), which is also mentioned in their *Limitations*. Therefore, the quality of stylized renderings is also affected. Our approach generates both precise geometry and stylization following the artistic style of the style reference.

Quantitative Results

Following the measurement in LSNV, we use a warped LPIPS metric (Zhang et al. 2018) to measure the consistency across different views. We utilize FlowNetS (Dosovitskiy et al. 2015) to compute the optical flow from a ground truth image I_x to another I_y. Subsequently, a warped mask M is generated based on the optical flow. Finally, we warp the corresponding stylized images $\hat{I}_x$ to $\hat{I}_y$ and calculate their distance along with M. The distance score is formulated as:

$$E(\hat{I}_x, \hat{I}_y) = LPIPS(M \odot Warp(\hat{I}_x, \hat{I}_y)), \tag{8}$$

where $\odot$ denotes element-wise multiplication.

We compare our method with baselines on LLFF dataset, reporting average warped distance score on 5 style references. We randomly choose 20 frame pairs $(\hat{I}_t, \hat{I}_{t+1})$ and $(\hat{I}_t, \hat{I}_{t+7})$ from each scene for short-range and long-range consistency respectively.

Ablation Study

Direct Image stylization on reconstruction appearance. We inverse sphere-to-cubemap retrieval to extract a cubemap showing the reconstruction appearance of the scene and use the cubemap as content input of the image stylization network instead of noise $\boldsymbol{z}$. We report the experimental results in Fig. 5, where we observe there is abrupt color in some areas impairing the stylization quality. We suggest that this is because the appearance is not uniformly distributed on the style pattern space. Thus, a small region in the cubamap might represent a wide area of the scene appearance, magnifying abrupt color in the appearance.

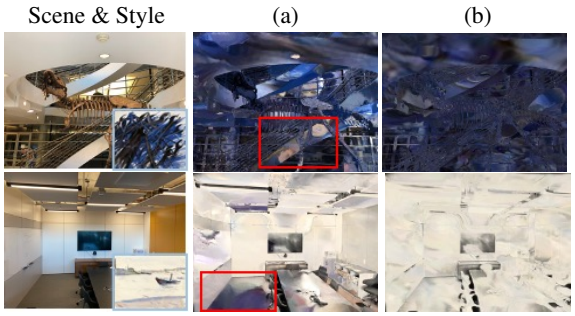

Figure 5: Ablation study on direct image stylization on reconstruction appearance. (a) The results of using reconstruction appearance cubemap as content input for image stylization. (b) The results of our method (using a noise image as content input and add a visual prompt in the bottleneck of image stylization network.)

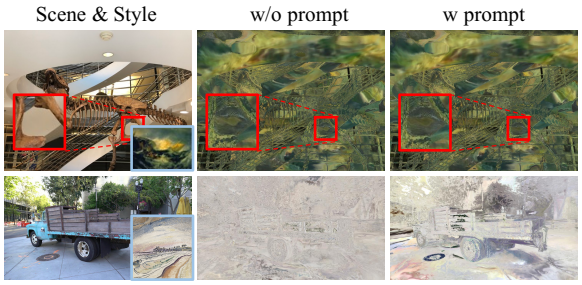

Figure 6: Ablation study for the visual prompt. The visual prompt alleviates the disorganized results and improve the visual quality.

With and without visual prompt. To investigate the effect of introducing a visual prompt, we evaluate the performance when it is removed. The result in Fig. 6 shows that directly using image style transfer network to generate style patterns can realize stylization, but the results are obviously disorganized without considering geometry information. It demonstrates that the visual prompt helps to adapt the style patterns to be aware of geometry information, thus the appearance of the scene can be stylized more harmoniously.

Conclusion

In this paper, we present a Prompt-based Neural Style Mapping (PNeSM) to transfer arbitrary styles to arbitrary 3D scenes. We take advantage of the powerful reconstruction capability of NeRF and completely disentangle appearance and geometry by mapping the appearance into a 2D style pattern space. By fusing the ability of texture information extraction in pre-trained 2D stylization network and effectiveness of prompt learning for fine-tuning, we achieve pleasant 3D scene stylization by stylizing the appearance of the scene in the 2D style pattern space. Extensive experimental results demonstrate the effectiveness and superiority of our method.

Acknowledgements

This work was supported in part by Zhejiang Province Program (2022C01222, 2023C03199, 2023C03201, 2019007, 2021009), the National Program of China (62172365, 2021YFF0900604, 19ZDA197), Ningbo Program(2022Z167), and MOE Frontier Science Center for Brain Science & Brain-Machine Integration (Zhejiang University).

References

Bahng, H.; Jahanian, A.; Sankaranarayanan, S.; and Isola, P. 2022. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 1(3): 4.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cao, X.; Wang, W.; Nagao, K.; and Nakamura, R. 2020. Psnet: A style transfer network for point cloud stylization on geometry and color. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer vision*, 3337–3345.

Chen, D.; Liao, J.; Yuan, L.; Yu, N.; and Hua, G. 2017. Coherent online video style transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, 1105–1114.

Chen, J.; Ji, B.; Zhang, Z.; Chu, T.; Zuo, Z.; Zhao, L.; Xing, W.; and Lu, D. 2023. TeSTNeRF: text-driven 3D style transfer via cross-modal learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 5788–5796.

Chiang, P.-Z.; Tsai, M.-S.; Tseng, H.-Y.; Lai, W.-S.; and Chiu, W.-C. 2022. Stylizing 3D scene via implicit representation and HyperNetwork. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1475–1484.

Cui, L.; Wu, Y.; Liu, J.; Yang, S.; and Zhang, Y. 2021. Template-Based Named Entity Recognition Using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1835–1845.

Deng, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; and Xu, C. 2021. Arbitrary video style transfer via multi-channel correlation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1210–1217.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; and Brox, T. 2015. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 2758–2766.

Fan, Z.; Jiang, Y.; Wang, P.; Gong, X.; Xu, D.; and Wang, Z. 2022. Unified implicit neural stylization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, 636–654. Springer.

Fridovich-Keil, S.; Yu, A.; Tancik, M.; Chen, Q.; Recht, B.; and Kanazawa, A. 2022. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5501–5510.

Gao, C.; Gu, D.; Zhang, F.; and Yu, Y. 2018. Reconet: Real-time coherent video style transfer network. In *Asian Conference on Computer Vision*, 637–653. Springer.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

Höllein, L.; Johnson, J.; and Nießner, M. 2022. Stylemesh: Style transfer for indoor 3d scene reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6198–6208.

Huang, H.-P.; Tseng, H.-Y.; Saini, S.; Singh, M.; and Yang, M.-H. 2021. Learning to stylize novel views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13869–13878.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Huang, Y.-H.; He, Y.; Yuan, Y.-J.; Lai, Y.-K.; and Gao, L. 2022. StylizedNeRF: consistent 3D scene stylization as stylized NeRF via 2D-3D mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18342–18352.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, 709–727. Springer.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 694–711. Springer.

Karayev, S.; Trentacoste, M.; Han, H.; Agarwala, A.; Darrell, T.; Hertzmann, A.; and Winnemoeller, H. 2013. Recognizing image style. *arXiv preprint arXiv:1311.3715*.

Knapitsch, A.; Park, J.; Zhou, Q.-Y.; and Koltun, V. 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4): 1–13.

Kopanas, G.; Philip, J.; Leimkühler, T.; and Drettakis, G. 2021. Point-Based Neural Rendering with Per-View Optimization. In *Computer Graphics Forum*, volume 40, 29–43. Wiley Online Library.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045–3059.

Li, C.; and Wand, M. 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, 702–716. Springer.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4582–4597.

Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; and Yang, M.-H. 2017. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30.

Liu, K.; Zhan, F.; Chen, Y.; Zhang, J.; Yu, Y.; El Saddik, A.; Lu, S.; and Xing, E. P. 2023. StyleRF: Zero-shot 3D Style Transfer of Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8338–8348.

Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Sun, Z.; Li, Q.; and Ding, E. 2021a. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6649–6658.

Liu, X.; Ji, K.; Fu, Y.; Du, Z.; Yang, Z.; and Tang, J. 2021b. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Mildenhall, B.; Srinivasan, P. P.; Ortiz-Cayon, R.; Kalantari, N. K.; Ramamoorthi, R.; Ng, R.; and Kar, A. 2019. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4): 1–14.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, 405–421. Springer.

Mu, F.; Wang, J.; Wu, Y.; and Li, Y. 2022. 3d photo stylization: Learning to generate stylized novel views from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16273–16282.

Nguyen-Phuoc, T.; Liu, F.; and Xiao, L. 2022. SNeRF: stylized neural implicit representations for 3D scenes. *ACM Transactions on Graphics (TOG)*, 41(4): 1–11.

Park, D. Y.; and Lee, K. H. 2019. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5880–5888.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473.

Sun, C.; Sun, M.; and Chen, H.-T. 2022. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5459–5469.

Ulyanov, D.; Lebedev, V.; Vedaldi, A.; and Lempitsky, V. 2016. Texture networks: Feed-forward synthesis of textures and stylized images.

Wang, W.; Yang, S.; Xu, J.; and Liu, J. 2020. Consistent video style transfer via relaxation and regularization. *IEEE Transactions on Image Processing*, 29: 9125–9139.

Xiang, F.; Xu, Z.; Hasan, M.; Hold-Geoffroy, Y.; Sunkavalli, K.; and Su, H. 2021. Neutex: Neural texture mapping for volumetric neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7119–7128.

Zhang, K.; Kolkin, N.; Bi, S.; Luan, F.; Xu, Z.; Shechtman, E.; and Snavely, N. 2022. ARF: Artistic Radiance Fields.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhong, Z.; Friedman, D.; and Chen, D. 2021. Factual Probing Is [MASK]: Learning vs. Learning to Recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5017–5033.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.